浅谈辽宁省互联网内容综合监管平台之互联网新闻出版舆情监测系统的应用

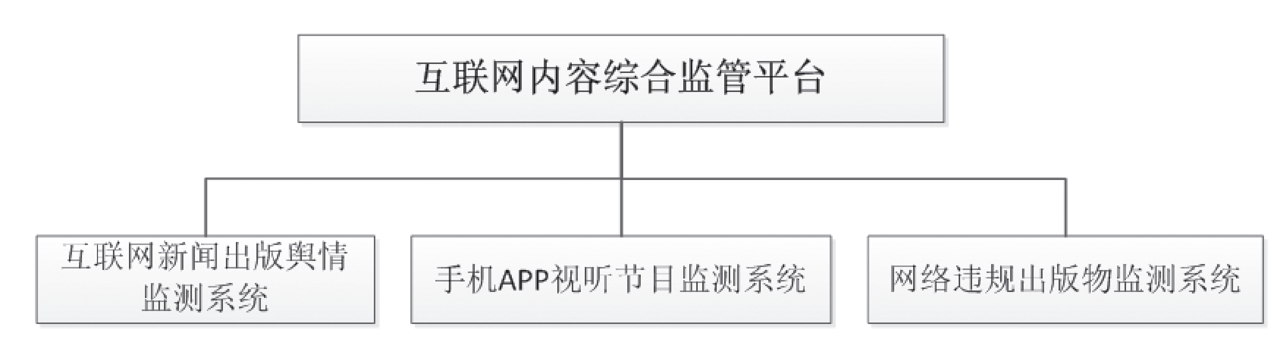
摘　要：本文阐述了互联网新闻出版舆情监测系统的数据采集技术分析。
关键词：系统；舆情主题聚焦爬虫；文本及情感分析技术
中图分类号：G206　　文献标识码：A
文章编号：1671-0134（2017）06-099-02　　DOI：10.19483/j.cnki.11-4653/n.2017.06.031

■文 / 刘　宁

1. 概述

随着互联网产业的飞速发展，互联网出版产业的发展空间也在不断拓展。但是，网络出版中存在的问题也必须引起我们的高度重视，通过本系统可以及时发现网上传播的有害网络出版物与负面舆情，并进行下载取证，全面、完整、详实地为检测部门提供日常监测数据和信息。

互联网内容综合监管平台是在统一的数据采集、数据分析、统计编报模块基础上，针对手机 APP 视听节目、互联网新闻出版舆情、网络违规出版物等不同监测领域，形成的一套可扩展的、一体化的智能监测综合平台。平台主要包括“互联网新闻出版舆情监测”“手机 APP 视听节目监测”“网络违规出版物监测”3 个组成部分，如下图所示。



可以全面监测互联网中关于新闻出版的实时舆情热点、舆情专题、手机 APP 软件中发布的视听节目以及各类网络出版物（例如网络文学、网络漫画、网络游戏等）在网络中传播情况，及时发现网上传播的违规视听节目、有害网络出版物与负面舆情，并进行下载取证，全面、完整、翔实地为监测部门提供日常监测数据和信息。

下面本文将着重介绍互联网内容综合监管平台中的互联网新闻出版舆情监测系统的数据采集。

2. 互联网新闻出版舆情监测系统

2.1 舆情监测系统中的数据处理

首先网络蜘蛛从互联网上抓取数据，一边抓取数据一边将已抓取的数据信息发送给应用服务器，应用服务器交给智能代理进行处理。

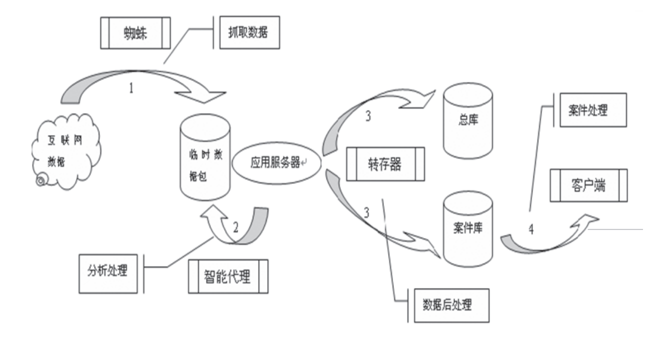
智能代理是系统中实现核心功能的子系统，对所有抓取的网络数据进行全面的分析过滤，识别出所监管的非法信息，提交给其他子系统做进一步处理。智能代理能够及时地自学习完善自己的知识体系，提高自身的智能性。

用户只需要设定要抓取站点的首页地址，蜘蛛程序就会按设定的站点下载相应的网页并传给后台处理程序做进一步的处理，并根据设定的更新周期，定期对各站点上新发布的网页或者更新了的网页进行及时抓取。

将抓取的数据打包成一个个临时数据包，然后将新数据任务发送给应用服务器，服务器选择一个空闲的智能代理，将这个任务分配给它进行分析处理，处理完后将这个处理任务反馈给应用服务器，服务器再将此任务分配给一个空闲的转存器。

转存器主要做一些处理工作，将系统发现的疑似案件数据及所有的临时数据存入到案件库和总库当中，通过客户端查看案件信息。根据发现案件的 URL 解析出其 IP 地址；将按规则分类出的案件按规则号对其文本内容标红；将案件从临时库转存到总库的案件库中；统计某个临时表发现的案件类型及其案件数并向服务器报警；将正常的信息都转存到总库的 Total 库中。

转存器再做进一步的处理，将有疑似违规的舆情信息导入到数据库，将正常的信息也导入到总库中，如果发现违规，则根据违规的类型通知负责监管这一主题的用户，客户端用户再对案件进行审计、反馈、确认、打印等功能。整个系统的数据处理流程如下图所示：



2.2 文本及情感分析技术

通过互联网各个信息系统传播的舆情事件信息，除了用于反应事件客观事实外，也表达了用户观点和情感，例如对该事件的支持、反对或中立态度。这些情感态度多数是通过互联网上的普通网民发表的文本信息表达出来，包含着人们对社会各种现象的不同观点和立场，个人和组织越来越多地把网络上的情感观点信息用于制定决策方面，从而使得情感分析技术应运而生。

情感分析技术对网络舆情事件发展走势的描述和预测有十分重要的作用，但是，由于网络舆情信息的多样性和中文文本处理的特殊性，针对网络舆情事件的中文情感分析面临诸多难点：

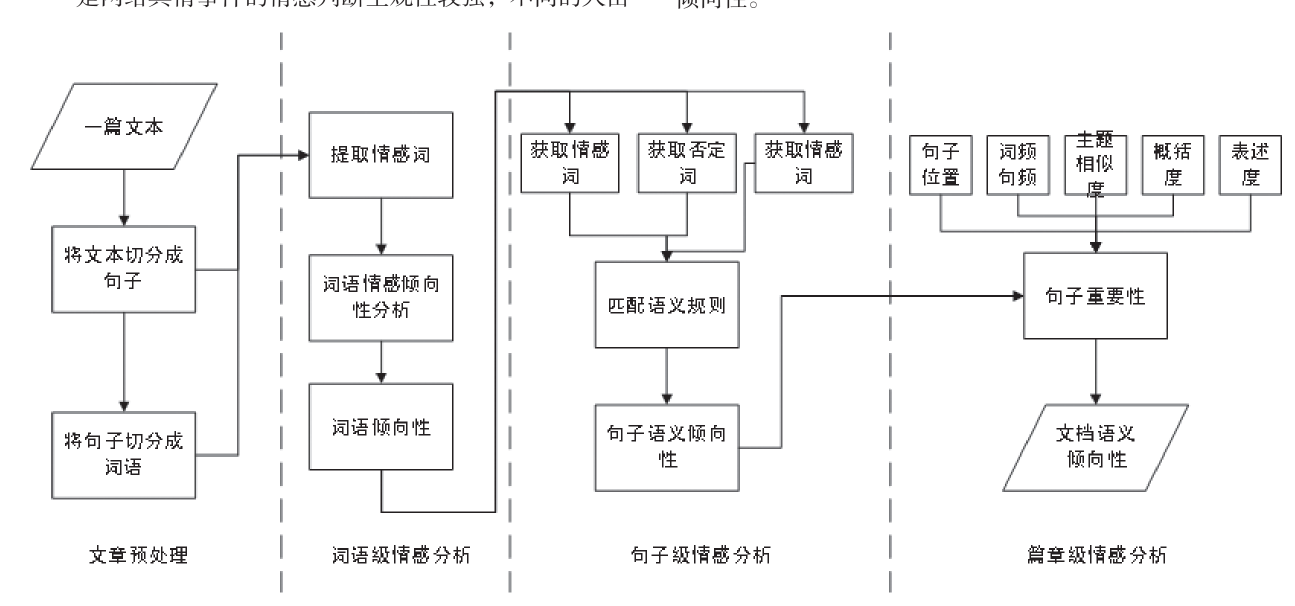
一是网络舆情事件的情感判断主观性较强，不同的人由于其身份背景、认知水平等限制，对同一信息的情感判断并不一致，因此其判定规则没有统一标准，因此由机器判定信息情感造成很大困难。

二是网络信息的载体较多，数据格式和类型不统一。网络舆情事件可以通过新闻、博客等长文本表述，又可以通过新型的论坛、微博等短文本进行传播。书面语言与口语混杂出现，新的网络词汇和变种词语大量增加，这种信息特征使得情感分析的难度大大增加。

三是网络舆情事件相关语料难以获取。目前互联网上舆情事件的相关中英文语料建设尚不完善，但情感分析所用的主要技术均需大量语料支撑。

四是中文情感分析难度较大。目前，对于英文的情感分析已做了很多研究，但中文由于其特殊性，准确度与中文分词、命名实体识别、句法分析等工具的准确度正相关。这些工具的准确度会大大影响中文情感识别的准确率。

下图为文本情感分析流程。首先输入一篇文本，进行文本的预处理，即将文本切分成句子，再将句子切分为词语。第二步进行词语级情感分析，得到每个句子中的词语情感倾向，第三步应用每个句子中词语的情感倾向进行句子级情感分析，获得每句话的情感倾向，最后，计算每句话在文章中的重要性，结合句子的情感倾向，最终输出该文档的正负面倾向性。



文本情感分析流程图

最终互联网新闻出版舆情分析系统可以实现对涉及全国、涉我（新闻出版相关的）的境内外热点、有害信息和涉稳的行动性信息进行主动发现，并对其传播进行追溯；支持业务相关的特定社会群体关注的热点的探测与发现；实现以热点云形式以及多热点分析指数来刻画网络热点。

通过对专题的分析，完成操作人员对特定关注主题，以及设置主题或事件为驱动的监控任务，实现对数据的主动采集、分析、统计到简报生成一站式服务，支持对事件走势情况、当前影响力情况、阶段演化分析、信息溯源跟踪、社交网络传播、网络推手识别、网民区域分布、网民情感分析与观点提炼、简报自动生成等功能。

（**作者单位：**辽宁省广播电视及信息网络视听节目传播监测中心 ）